PAPER PRESENTATION

VeriTrace: A Multitask Disaster Dataset for Real–Fake Verification and Source Tracing

Surbhi Madan[1], Yusuke Yasuda[1] and Junichi Yamagishi[1]

National Institute of Informatics, Tokyo[1]

DOI: [10.1109/OJSP.2026.3684389]

IEEE Open Journal of Signal Processing · ICIP 2026 Spotlight Presentation

OUTLINE

01 Introduction

02 Research Gap

03 Building VeriTrace

04 Experimental Setup

05 Results & Discussion

06 Conclusion

Fake disaster images now spread as fast as real ones

Social media spreads disaster information fast — and just as fast, **disaster deepfakes.**

> **2022 Shizuoka Typhoon, Japan**
> AI-generated flood images spread widely on social media, severe enough to prompt an official government response.

Why disaster deepfakes are hard to detect:

- Semantically complex scenes : chaotic rubble, fire, flood reflections
- Scarce labeled training data for this domain
- New, unseen generative models appear constantly



FIGURE: Semantic correspondence between real disaster images and top five synthetic samples.

RESEARCH GAP

Existing datasets don't cover disaster deepfakes

✗

General deepfake sets

FaceForensics++ [9], DFDC [10], GenImage [12], WildFake [13], Community Forensics [14] — built for faces or generic scenes, not disaster semantics.

⚠

Incidents1M

Large real-world disaster imagery [3] —but no paired synthetic (fake) counterparts for forensic training.

✗

Source-tracing studies

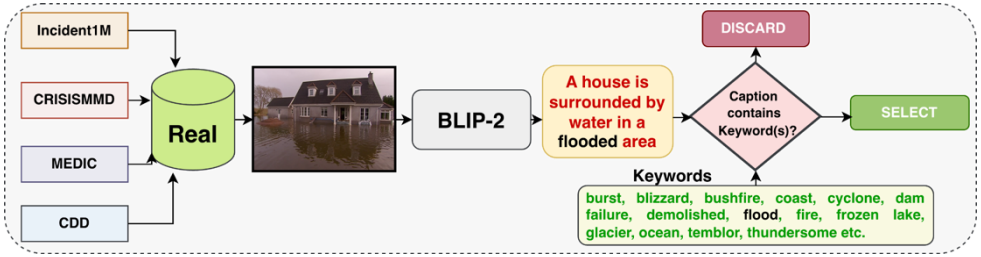
WILD [8], MADE [5], DeSI [20] test attribution across only ~10–20 generators — far from real-world scale.

The gap: no large-scale, multitask dataset exists that pairs real disaster images with fakes and labels them for *both* detection and source attribution, at a scale of hundreds–thousands of generators.

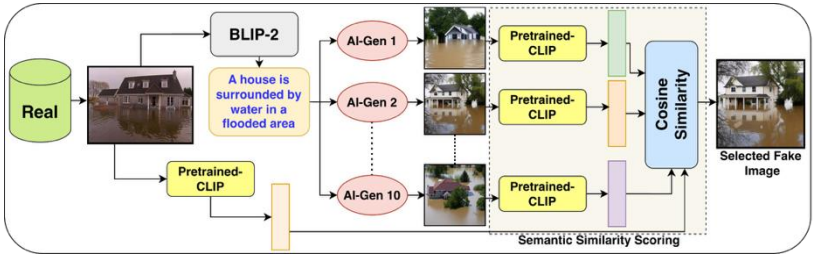
DATASET CONSTRUCTION PIPELINE

Five steps, two figures — I will take them one at a time

1 Collect real images 2 Caption 3 Keyword filter 4 Generate 5 Rank & select



Steps 1–3 collect · caption · filter



Steps 4–5 generate · rank and select

1 Aggregate real disaster images from four public datasets.

2 Caption each image with BLIP-2.

3 Keep it only if the caption names a disaster.

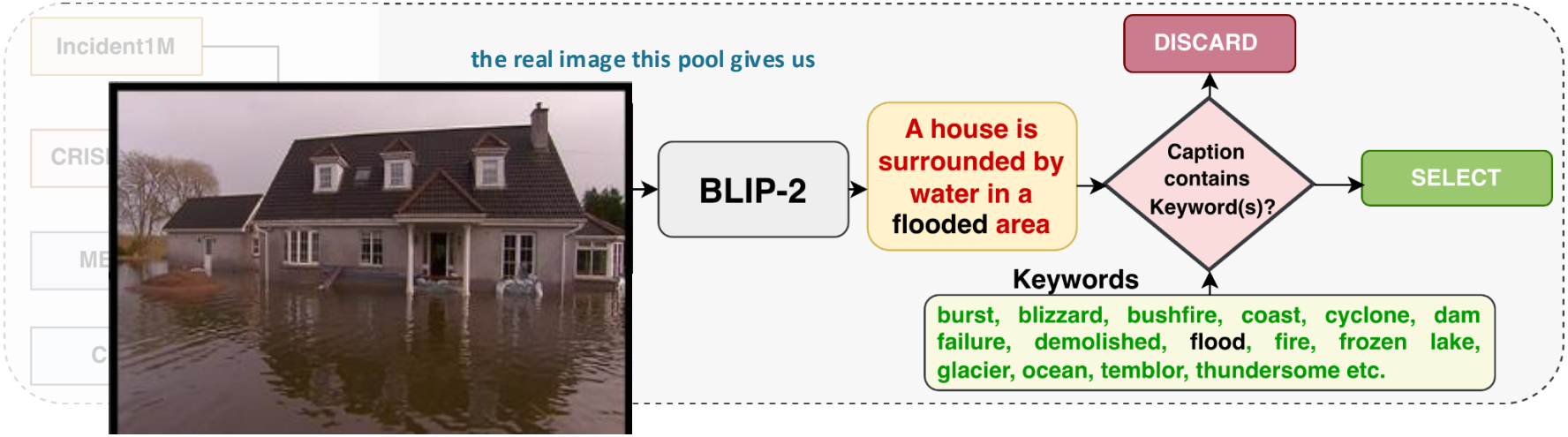
4 Feed the caption to generators picked at random.

5 Rank the candidates by CLIP similarity, keep the best.

DATASET CONSTRUCTION PIPELINE

Four public datasets, pooled into one set of real images

1 **Collect real images** 2 Caption 3 Keyword filter 4 Generate 5 Rank & select

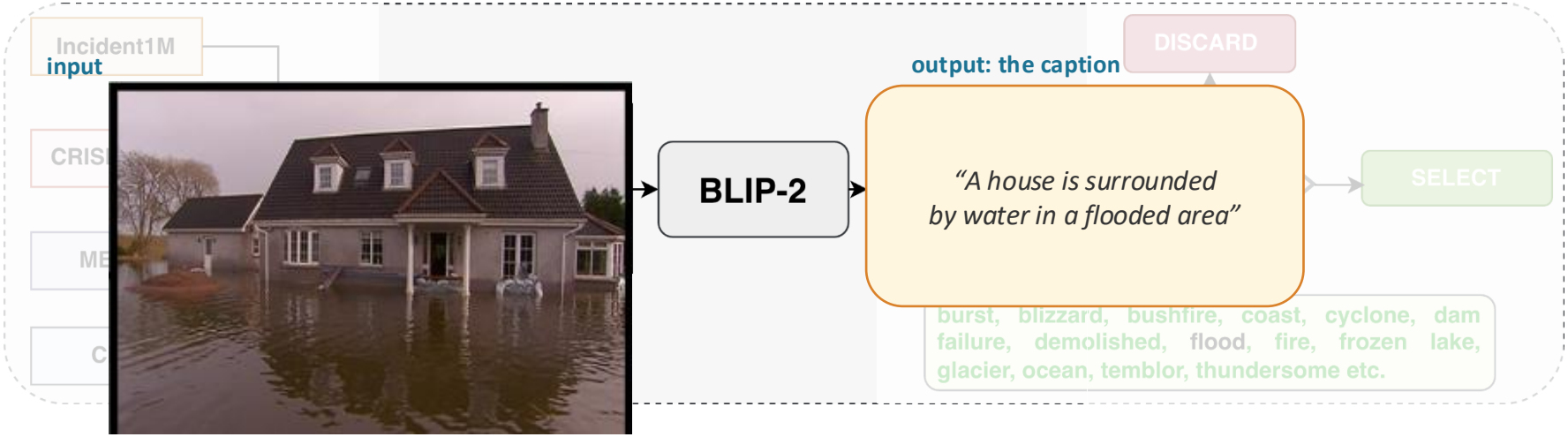


Incidents1M, CRISIS, CDD and MEDIC are pooled into a single set of real disaster images. Everything that follows starts from one of them.

DATASET CONSTRUCTION PIPELINE

A caption turns the picture into a prompt

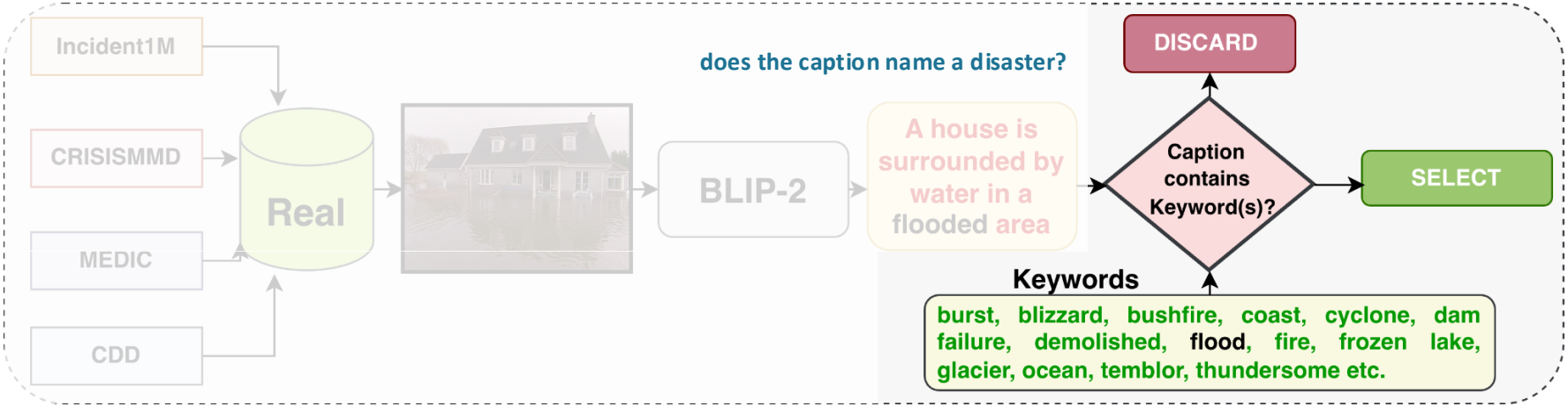
1 Collect real images 2 **Caption** 3 Keyword filter 4 Generate 5 Rank & select



DATASET CONSTRUCTION PIPELINE

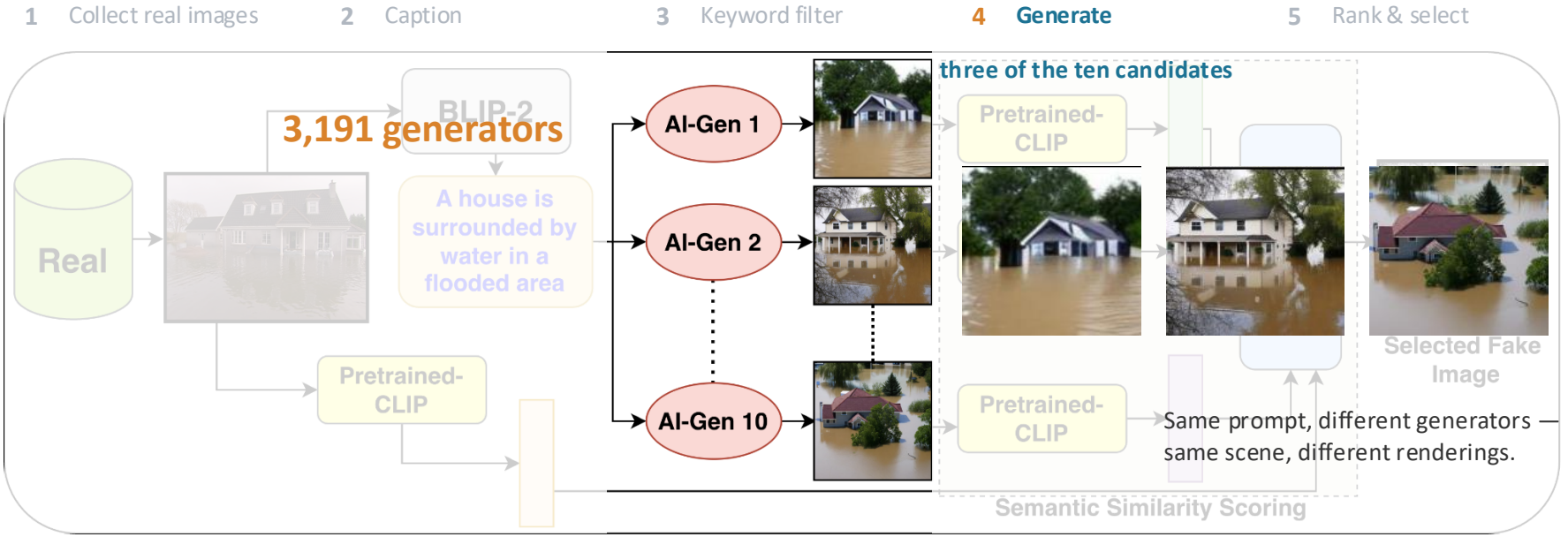
Only captions that name a disaster survive

1 Collect real images 2 Caption **3 Keyword filter** 4 Generate 5 Rank & select



DATASET CONSTRUCTION PIPELINE

One prompt, many generators, many candidates

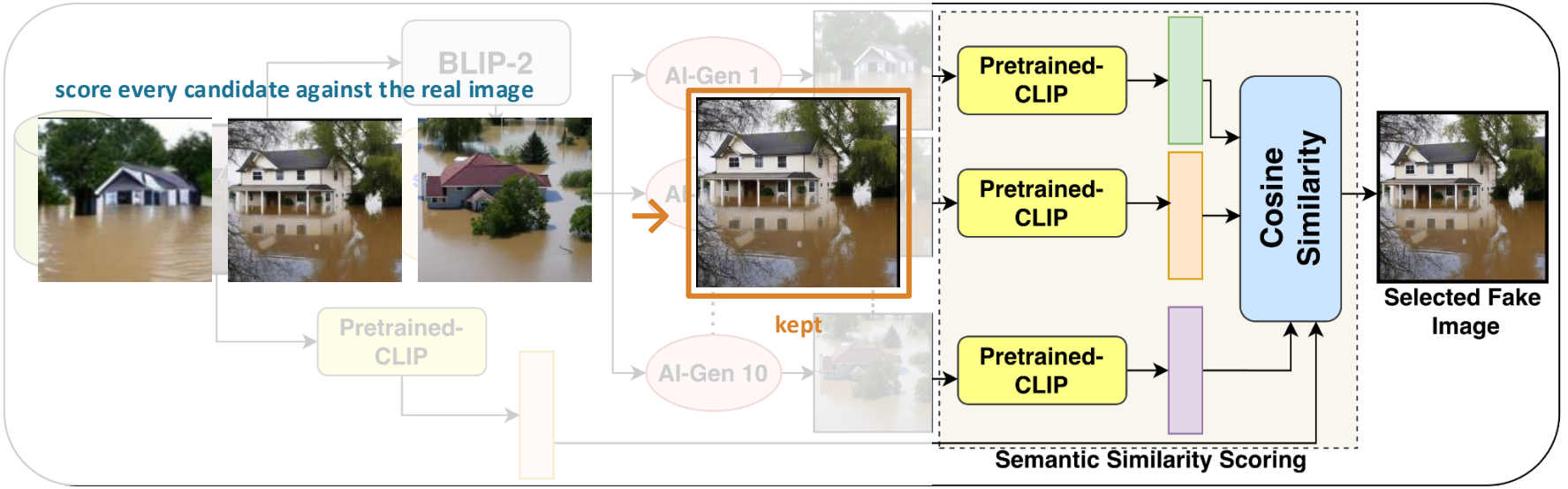


screened from 7,000+ models on Hugging Face, following the Community Forensics [14] approach. Ten are drawn at random for each real image.

DATASET CONSTRUCTION PIPELINE

Keep the candidate closest to the real image

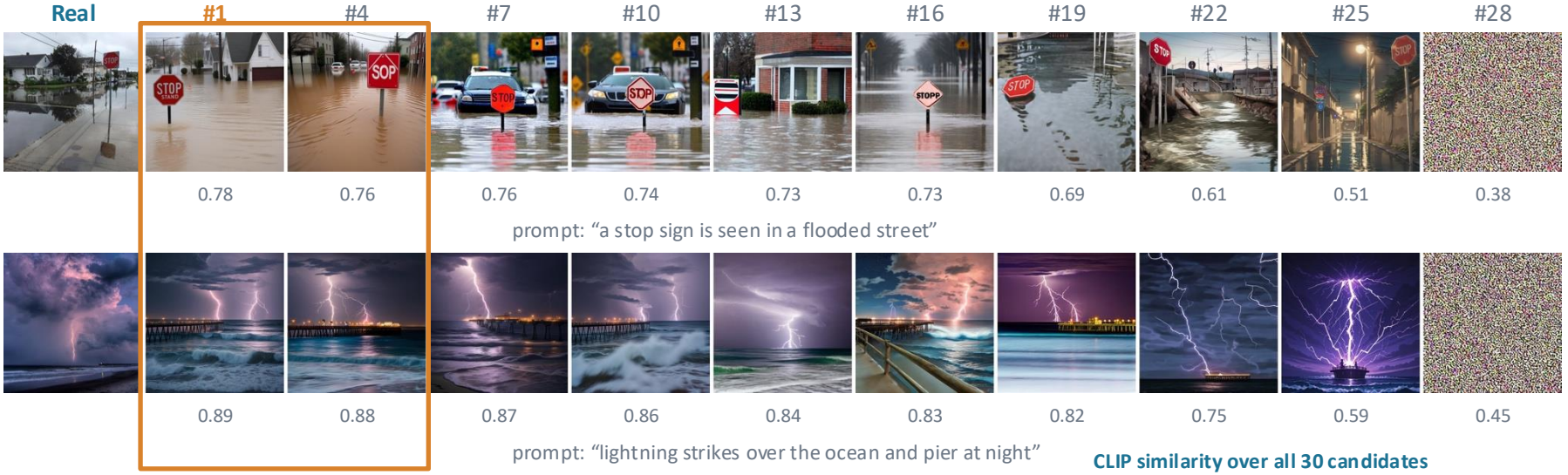
1 Collect real images 2 Caption 3 Keyword filter 4 Generate 5 Rank & select



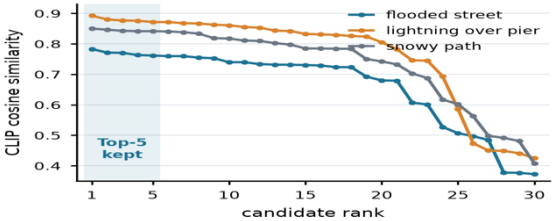
CLIP cosine similarity between the real image and each candidate. The highest score becomes the paired fake.

WHY TOP-N SELECTION MATTERS

Go deeper down the ranking and the candidates stop being the disaster scene



CLIP similarity over all 30 candidates



By rank 19 the flooded street is a dry alley, and by rank 28 it is noise.

The pier survives longer, but rank 25 is an abstract flash and rank 28 is noise.

Ten of thirty candidates shown. Keeping the top five keeps the scene.

GENERATORS: OPEN-WEIGHT AND COMMERCIAL

3,191 open-weight models and 10 commercial APIs

Open group = open-weight models

- Collected through the diffusers library [27][28], which gives one standard interface
- 7,000+ candidates screened on Hugging Face → 3,191 that ran end to end
- Used for both tasks; 902 of them qualify for source tracing

Closed group = commercial APIs

- DALL·E 2 [29]
- DALL·E 3 [30]
- Ideogram [31]
- Midjourney [32]
- Leonardo AI [33]
- FLUX.1 Dev / Schnell [34]
- SD XL-Turbo [35]
- SD 3.5 Large [36]
- Gemini 2.5 Flash Image [37]

None of the ten appear in any detector's training data — our out-of-distribution test.

What it cost to build

~15 days
screening 3,191 generators

40+ days
generation on 4–8 GPUs (40 GB)

~15 days / $1,500
commercial API generation

VERITRACE: MULTITASK DATASET

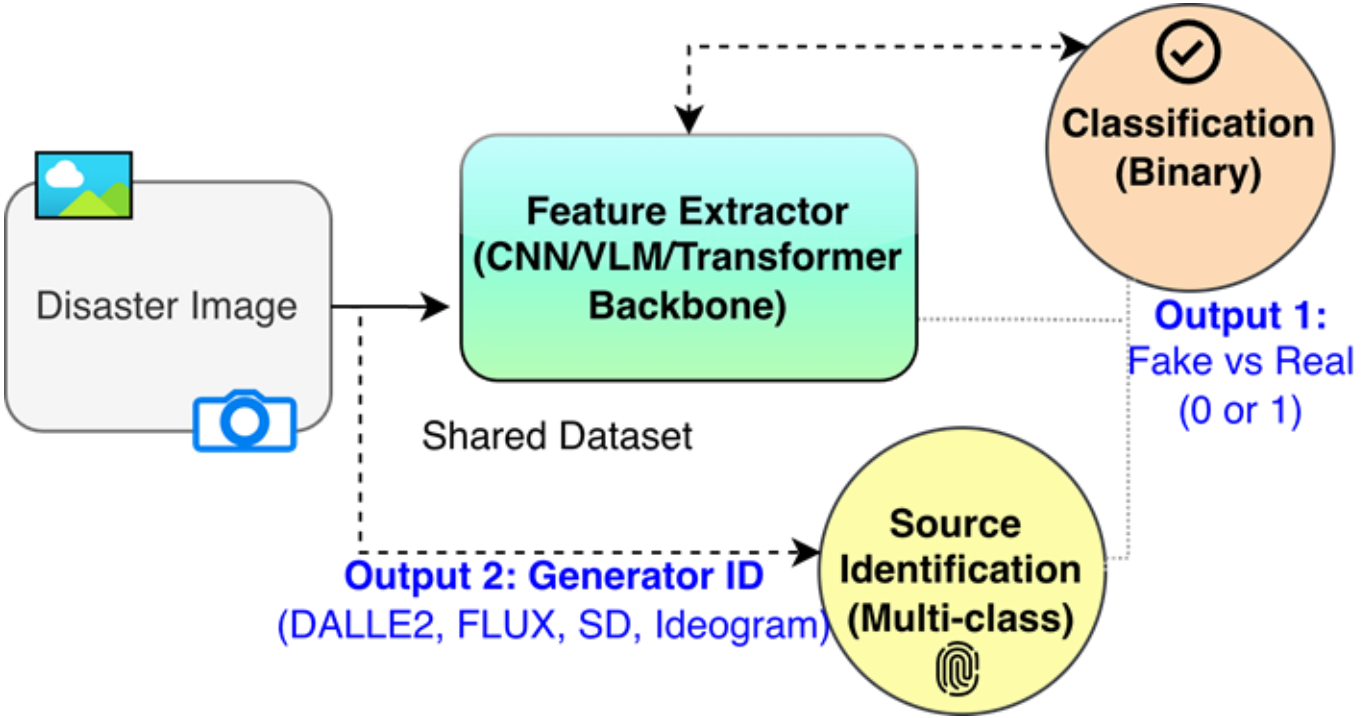


FIGURE: Proposed VeriTrace Dataset Schema Supporting Joint Learning of Deepfake Detection and Source Tracing in Disaster-Related Images.

DATASET STATISTICS

One dataset, two tasks — and two different generator counts

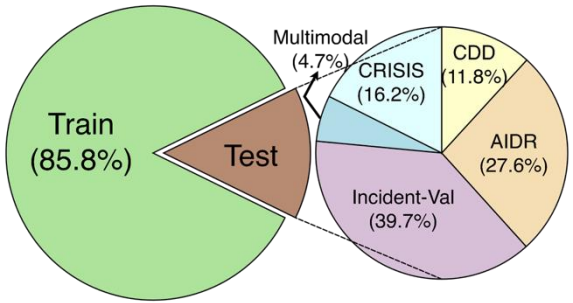
Binary detection

3,191 generators (+ 10 commercial)

Train: 208,209 real + 208,209 fake

Test: Incident Val 13,716 · AIDR 9,545 ·

CRISIS 5,598 · CDD 4,077 · Multimodal 1,643



(a) Train / test composition across the source datasets.

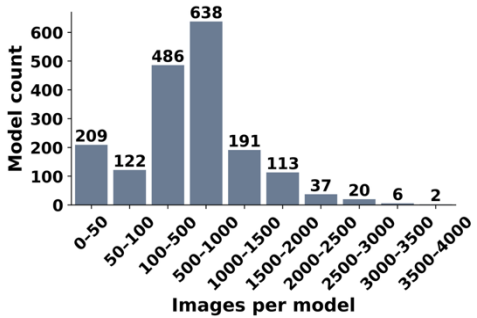
Source tracing

912 generators (902 open-weight + 10 commercial)

Train: 456,000 images

Test: Open 22,250 / 89 · Closed 10,000 / 10 ·

Combined 32,250 / 99



(b) Images per generator is very uneven. Keeping only models with ≥500 images takes 1,825 → 912 — this is why source tracing uses fewer generators.

EXPERIMENTAL SETUP

All three detectors are evaluated zero-shot — no fine-tuning on VeriTrace

Hybrid

AIDE [40]

Hybrid semantic + frequency features (CLIP embeddings + patch-wise frequency analysis)

Vision

Community Forensics [14]

Pure vision baseline — ViT-S / ConvNeXt-S, evaluated at 224^2 and 384^2

Reasoning

ThinkFake

Reasoning-driven MLLM for explainable forgery detection

Evaluation Metrics

Detection

Average Accuracy & Average Precision (AP)

Source Tracing

Weighted Multiclass Accuracy & Weighted Multiclass AP

DETECTION RESULTS: OPEN-WEIGHT GENERATORS

The vision baseline works here — the other two do not

TABLE 1 (open set): balanced Real vs. Top-1. Each cell is Accuracy / AP; ThinkFake reports accuracy only.

Dataset	#Samples	AIDE-ProGAN	AIDE-SD v1.4	AIDE-GenImg	CF-224	CF-384	ThinkFake
Incident Val	**13,716**	**0.49 / 0.52**	**0.58 / 0.78**	**0.54 / 0.68**	**0.89 / 0.97**	**0.94 / 0.99**	**0.57**
Multimodal	1,643	0.53 / 0.69	0.50 / 0.46	0.56 / 0.77	0.90 / 0.98	0.98 / 1.00	0.66
CDD	4,077	0.55 / 0.72	0.50 / 0.49	0.55 / 0.76	0.84 / 0.94	0.86 / 0.98	0.68
AIDR	9,545	0.54 / 0.75	0.51 / 0.72	0.53 / 0.77	0.87 / 0.96	0.86 / 0.98	0.66
CRISIS	5,598	0.54 / 0.73	0.51 / 0.59	0.57 / 0.84	0.91 / 0.98	0.95 / 0.99	0.61

- CF-384 reaches 0.94 / 0.99 on Incident Val — a standard vision model, used zero-shot, copes here.
- AIDE stays near 0.50 accuracy and ThinkFake near 0.60: two of three approaches already struggle.

DETECTION RESULTS: COMMERCIAL GENERATORS

The same models collapse — accuracy falls, AP does not

TABLE 1 (closed group): the identical detectors on images from the ten commercial APIs.

Dataset	#Samples	AIDE-ProGAN	AIDE-SD v1.4	AIDE-GenImg	CF-224	CF-384	ThinkFake
Incident Val	**13,716**	**0.50 / 0.50**	**0.50 / 0.50**	**0.50 / 0.50**	**0.34 / 0.37**	**0.30 / 0.36**	**0.58**
Multimodal	1,643	0.63 / 0.70	0.58 / 0.78	0.64 / 0.82	0.50 / 0.97	0.50 / 0.98	0.69
CDD	4,077	0.60 / 0.68	0.57 / 0.78	0.62 / 0.82	0.60 / 0.72	0.61 / 0.89	0.70
AIDR	9,545	0.65 / 0.77	0.58 / 0.83	0.63 / 0.85	0.64 / 0.83	0.66 / 0.92	0.71
CRISIS	5,598	0.64 / 0.77	0.59 / 0.84	0.63 / 0.85	0.65 / 0.86	0.76 / 0.96	0.67

- CF-384 on Incident Val: 0.94 → 0.30 accuracy, below chance. AIDE sits at exactly 0.50 / 0.50.
- On Multimodal, CF-384 keeps 0.98 AP with 0.50 accuracy: the ranking survives, the threshold does not.

CORRECT AND INCORRECT EXAMPLES

All eight images are fake — the bottom row is what a missed detection looks like



open-weight generators

Ground truth: Fake

Predicted: Fake

→ correct

closed-group generators

Ground truth: Fake

Predicted: Real

→ missed

FIGURE: Fake images from open-weight generators, correctly detected (top), and from closed-group generators, missed (bottom).

GRAD-CAM:
WHERE THE DETECTOR LOOKS

Right region → correct prediction. Background → wrong prediction.

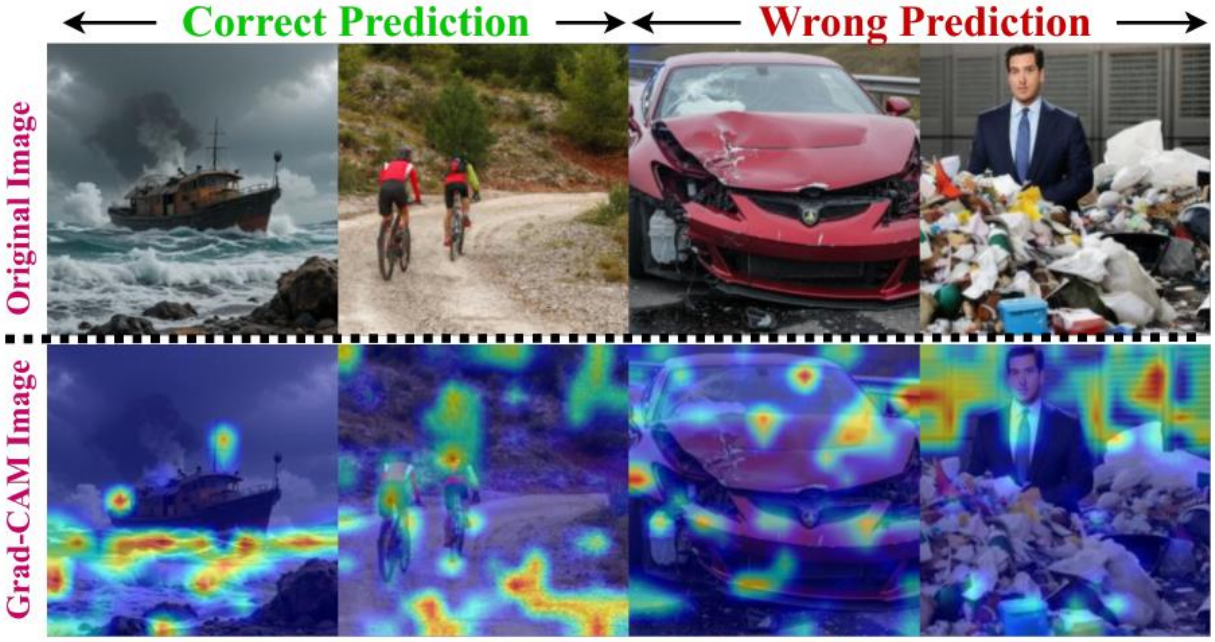


FIGURE 7: Grad-CAM for correct (left) and incorrect (right) predictions. Top: originals. Bottom: Grad-CAM heatmaps.

SOURCE TRACING: WHICH OF 900+ GENERATORS?

Multi-class classification, not real vs. fake — and here the commercial models are the easiest to identify.

TABLE: Weighted source tracing performance on VeriTrace-ST for Open, Closed, and Combined groups. N denotes the number of unique generators.

		Sources	AIDE-ProGAN		AIDE-SD v1.4		AIDE-GenImg		CF-224		CF-384	
	Dataset	(N)	W-Acc	W-AP	W-Acc	W-AP	W-Acc	W-AP	W-Acc	W-AP	W-Acc	W-AP
Open Group	AIDR	18	0.15	0.61	0.12	0.56	0.13	0.55	0.15	0.59	0.26	0.72
	CDD	07	0.22	0.85	0.22	0.81	0.20	0.80	0.27	0.81	0.36	0.85
	CRISIS	27	0.07	0.33	0.07	0.32	0.07	0.29	0.13	0.32	0.20	0.45
	Incident_Val	37	0.11	0.36	0.09	0.32	0.08	0.31	0.14	0.36	0.21	0.46
Closed Group	AIDR	10	0.63	0.89	0.64	0.86	0.69	0.92	0.83	0.93	0.90	0.97
	CDD	10	0.62	0.84	0.57	0.80	0.67	0.86	0.78	0.91	0.87	0.95
	CRISIS	10	0.45	0.74	0.47	0.72	0.51	0.77	0.64	0.82	0.75	0.88
	Incident_Val	10	0.51	0.83	0.50	0.76	0.59	0.84	0.73	0.89	0.84	0.94
Combined	AIDR	28	0.32	0.69	0.31	0.64	0.33	0.66	0.39	0.69	0.49	0.81
	CDD	17	0.45	0.80	0.43	0.75	0.47	0.81	0.57	0.83	0.67	0.91
	CRISIS	37	0.18	0.41	0.18	0.39	0.19	0.39	0.27	0.44	0.35	0.56
	Incident_Val	47	0.19	0.45	0.18	0.41	0.18	0.41	0.26	0.46	0.34	0.56

BROADER IMPACT & RELEASE

A defensive resource, released through NII IDR

Responsible use

- Purpose is defensive: fact-checking, and helping emergency teams check images before acting on them.
- Added misuse risk is limited — the generators themselves are already public.
- Graphic content removed by automatic filtering followed by manual checking.

Dataset release

- Published through NII IDR, our institute's research data repository.
- Accepted by the repository; the terms of use are being finalised.
- Access by application, reviewed by the repository.
- Set up for research use in Japan first; requests from other countries are reviewed individually.

Please talk to us if you would like to use VeriTrace.

CONCLUSION

Disaster deepfakes are significantly harder to detect than human-centric ones

and VeriTrace is the first benchmark to test source tracing at a scale of 900+ generators.

Limitations

- **Text-to-image models only** — no GAN/autoencoder architectures
- **Static images only**, no video support
- Tracing across very large generator sets **remains open**

Future Work

- Integrate **artifact-aware + semantic** reasoning representations
- **Broaden generator diversity** for robustness
- Extend to **editing / compositional** forgery types

REFERENCES

Numbering follows the paper.

[3] E. Weber et al., “Incidents1M: a large-scale dataset of images with natural disasters, damage, and incidents,” IEEE TPAMI, 2022.

[5] S. Bonechi et al., “Who made this? Fake detection and source attribution with diffusion features,” arXiv:2510.27602, 2025.

[8] P. Bongini et al., “WILD: a new in-the-wild image linkage dataset for synthetic image attribution,” IJCNN, 2025.

[9] A. Rössler et al., “FaceForensics++: Learning to detect manipulated facial images,” ICCV, 2019.

[10] B. Dolhansky et al., “The DeepFake Detection Challenge (DFDC) dataset,” arXiv:2006.07397, 2020.

[12] M. Zhu et al., “GenImage: A million-scale benchmark for detecting AI-generated image,” NeurIPS, 2023.

[13] Y. Hong et al., “WildFake: A large-scale and hierarchical dataset for AI-generated images detection,” AAAI, 2025.

[14] J. Park and A. Owens, “Community Forensics: Using thousands of generators to train fake image detectors,” CVPR, 2025.

[18] J. Li et al., “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and LLMs,” ICML, 2023.

[20] K. Narayan et al., “DeSI: Deepfake source identifier for social media,” CVPR, 2022.

[24] F. Alam et al., “CrisisMMD: Multimodal Twitter datasets from natural disasters,” ICWSM, 2018.

[25] A. Verma, “Disaster Images Dataset,” Kaggle, 2021.

[26] F. Alam et al., “MEDIC: a multi-task learning dataset for disaster image classification,” Neural Comput. Appl., 2023.

[27][28] P. von Platen et al., “Diffusers: State-of-the-art diffusion models.”

[29]–[37] Commercial generators: DALL·E 2/3, Ideogram, Midjourney, Leonardo AI, FLUX.1, Stable Diffusion, Gemini 2.5 Flash Image.

[40] S. Yan et al., “A sanity check for AI-generated image detection (AIDE),” ICLR, 2025.



Yusuke Yasuda

yasuda@nii.ac.jp

National Institute of Informatics, Tokyo

Scan for more details

EXPERIMENTAL RESULTS

On a realistic mix, accuracy collapses while AP stays high

TABLE 2: Imbalanced Real vs. All setting — each real image is paired with all five of its fakes. Cells are Acc / AP; ThinkFake reports accuracy only. Overall rows per dataset.

Group	Dataset	Real / Fake	AIDE-ProGAN	AIDE-SD v1.4	AIDE-GenImg	CF-224	CF-384	ThinkFake
Open Set	Incident Val	13,716 / 68,580	0.18 / 0.85	0.31 / 0.94	0.26 / 0.91	0.85 / 0.99	0.90 / 0.99	0.55
Open Set	Multimodal	1,643 / 8,215	0.26 / 0.92	0.19 / 0.84	0.26 / 0.95	0.86 / 0.99	0.94 / 1.00	0.64
Open Set	CDD	4,077 / 20,385	0.26 / 0.92	0.18 / 0.83	0.26 / 0.94	0.79 / 0.98	0.80 / 1.00	0.68
Open Set	AIDR	9,545 / 47,725	0.26 / 0.94	0.18 / 0.86	0.23 / 0.94	0.81 / 0.99	0.78 / 1.00	0.64
Open Set	CRISIS	5,598 / 27,990	0.25 / 0.93	0.20 / 0.88	0.29 / 0.96	0.85 / 0.99	0.89 / 0.99	0.54
Closed Group	Incident Val	13,716 / 68,580	0.42 / 0.81	0.38 / 0.80	0.45 / 0.82	0.41 / 0.75	0.37 / 0.71	0.82
Closed Group	Multimodal	1,643 / 8,215	0.48 / 0.92	0.40 / 0.96	0.52 / 0.96	0.50 / 0.96	0.50 / 0.98	0.83
Closed Group	CDD	4,077 / 20,385	0.41 / 0.90	0.37 / 0.94	0.48 / 0.96	0.43 / 0.93	0.40 / 0.98	0.84
Closed Group	AIDR	9,545 / 47,725	0.49 / 0.94	0.40 / 0.96	0.50 / 0.98	0.52 / 0.97	0.49 / 0.98	0.82
Closed Group	CRISIS	5,598 / 27,990	0.50 / 0.94	0.42 / 0.97	0.50 / 0.96	0.53 / 0.98	0.67 / 0.93	0.79

AIDE accuracy falls to 0.18–0.30 on the open set while AP stays at 0.83–0.96: the ranking survives, the fixed threshold does not.

DATASET CONSTRUCTION PIPELINE

1 — Aggregate real images
Incidents1M, CRISIS, CDD, MEDIC (AIDR + Multimodal)

2 — Caption + filter
BLIP-2 captions validated against disaster keywords

3 — Generate candidates
10 sampled generators produce fakes per real image (of 3,191 total)

4 — CLIP similarity ranking
Embed real + candidates → cosine similarity → rank Top-1–Top-5

5 — Select best match
Highest-similarity fake retained as the paired counterpart

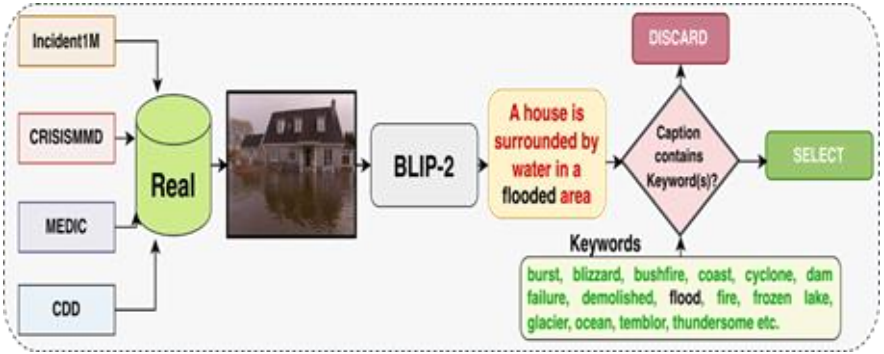


FIGURE: Automated Keyword Filtering for Constructing a VeriTrace Dataset with Authenticity and Source Labels.

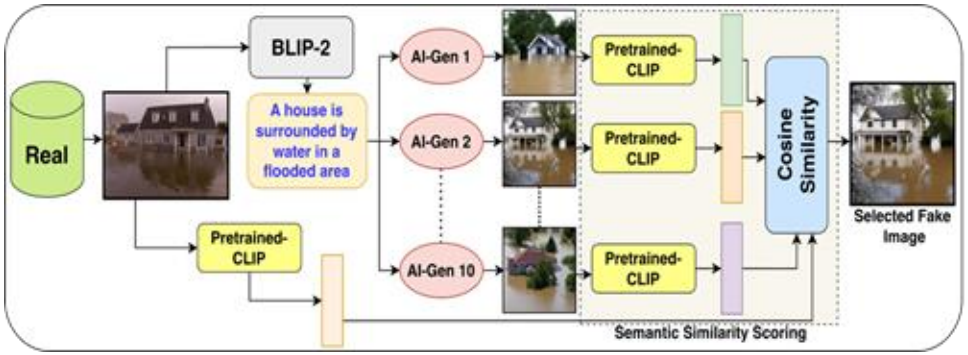


FIGURE: Dataset Generation Pipeline: Integrating BLIP-2 Captioning and Cosine Similarity for High-Fidelity Synthetic Image Selection.